International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 4, Issue 2, Apr 2014, 109-114

TRANS STELLAR
• Journal Publications • Research Consultancy

WEB-SCALE INFORMATION EXTRACTION FROM UNSTRUCTURED AND UNGRAMMATICAL DATA SOURCES

MADHAVI K. SARJARE[1] & S. L. VAIKOLE[2]

[1]Department of Computers, Yadavrao Tasgoankar College of Engineering and Management, Karjat, Mumbai, Maharashtra, India

[2]Department of Computers, Datta Meghe College of Engineering, Airoli, Navi Mumbai, Maharashtra, India

ABSTRACT

Information Extraction (IE) is the task of automatically extracting knowledge from text. The massive body of text now available on the World Wide Web presents an unprecedented opportunity for information extraction. However, information extraction on the Web is challenging due to the enormous variety of distinct concepts and structured expressed. The explosive growth and popularity of the worldwide web has resulted in a huge amount of information sources on the Internet. However, due to the heterogeneity and the lack of structure of Web information sources, access to this huge collection of information has been limited to browsing and searching.

Information extraction from unstructured and ungrammatical text on the Web, such as classified Ads, Auction listings, and web postings forums. Since the data is unstructured and ungrammatical, this information extraction precludes the use of rule-based methods that rely on consistent structures within the text or natural language processing techniques that rely on grammar. Posts are full of useful information, as defined by the attributes that compose the entity within the post.

Currently accessing the data within posts does not go much beyond keyword search. This is precisely because the ungrammatical and unstructured nature of posts makes extraction difficult, so the attributes remain embedded within the posts. These data sources are ungrammatical, since they do not conform to the proper rules of written language. Therefore, Natural Language Processing (NLP) based information extraction techniques are not appropriate.

As more and more information comes online, the ability to process and understand this information becomes more and more crucial. Data integration attacks this problem by letting users query heterogeneous data sources within a unified query framework, combining the results to ease understanding. However, while data integration can integrate data from structured sources such as databases, semi-structured sources such as that extracted from Web pages, and even Web Services, this leaves out a large class of useful information: unstructured and ungrammatical data sources.

We proposed a system based Machine Learning technique to obtain the structured data records from different unstructured and non-template based websites. The proposed approach will be implemented by collection of known entities and their attributes, which refer as "*reference set*," A reference set can be constructed from structured sources, such as databases, or scraped from semi-structured sources such as collections of Web pages. A reference set can even be constructed automatically from the unstructured, ungrammatical text itself. This project implements methods to exploit reference sets for extraction using machine learning techniques. The machine learning approach provides higher accuracy extractions and deals with ambiguous extractions, although at the cost of requiring human effort to label training data.

KEYWORDS: Natural Language Processing, Reference Set, Nested String List, Hypertrees

INTRODUCTION

The Internet provides access to numerous sources of useful information in textual form. Recently, there has been much interest in building systems that gather such information on a user's behalf. But because people format these information resources for use, mechanically extracting their content is difficult. Systems using such resources typically use hand-coded *wrappers*, customized procedures for information extraction.

Information extraction from unstructured, ungrammatical text on the Web such as classified ads, auction listings, and forum postings is a challenging work. Since the data is unstructured and ungrammatical, this information extraction precludes the use of rule-based methods that rely on consistent structures within the text or natural language processing techniques that rely on grammar.

Posts data are full of useful information, as defined by the attributes that compose the entity within the post. For example, consider the posts about cars from the online classified service. Each used car for sale is composed of attributes that define this car; and if we could access the individual attributes we could include such sources in data integration systems, and answer interesting queries. Such a query might require combining the structured database of safety ratings with the posts of the classified ads and the car review websites.

However, currently accessing the data within posts does not go much beyond keyword search. This is precisely because the ungrammatical, unstructured nature of posts makes extraction difficult, so the attributes remain embedded within the posts. These data sources are ungrammatical, since they do not conform to the proper rules of written language. Therefore, Natural Language Processing (NLP) based information extraction techniques are not appropriate. Further, the posts are unstructured since the structure can vary vastly between each listing. So, wrapper based extraction techniques will not work either. Even if one can extract the data from within posts, you would need to assure that the extracted values map to the same value for accurate querying.

LITERATURE SURVEY

Existing Systems

The combination of various input documents and variation of extraction targets causes different degrees of task difficulties. Since various IE systems are designed for various IE tasks, it is not fair to compare them directly. However, analyzing what task an IE system targets and how it accomplishes the task, can be used to evaluate this system and possibly extend to other task domains.

TSIMMIS: is one of the first approaches that give a framework for manual building of Web wrappers [7]. The main component of this project is a wrapper that takes as input a specification file that declaratively states (by a sequence of commands given by programmers) where the data of interest is located on the pages and how the data should be "packaged" into objects. Each command is of the form: [*variables*, *source*, *pattern*], where *source* specifies the input text to be considered, *pattern* specifies how to find the text of interest within the source, and *variables* are a list of variables that hold the extracted results. The special symbol '*' in a pattern means discard, and '#' means save in the variables.

Web OQL: is a functional language that can be used as query language for the Web, for semi structured data and for website restructuring [6]. The main data structure provided by Web OQL is the *hyper tree*. Hyper trees are arc-labeled

ordered trees which can be used to model a relational table, a Bib tex file, a directory hierarchy, etc. The abstraction level of the data model is suitable to support collections, nesting, and ordering.

W4F: (Wysiwyg Web Wrapper Factory) is a Java toolkit to generate Web wrappers [8]. The wrapper development process consists of three independent layers: *retrieval*, *extraction* and *mapping* layers. In the retrieval layer, a to-be processed document is retrieved (from the Web through HTTP protocol), cleaned and then fed to an HTML parser that constructs a parse tree following the Document Object Model (DOM). In the extraction layer, extraction rules are applied on the parse tree to extract information and then store them into the W4F internal format called Nested String List (NSL).

This project objective is to exploit reference sets for extraction using both automatic techniques and machine learning techniques. The automatic technique provides a scalable and accurate approach to extraction from unstructured, ungrammatical text. The machine learning approach provides even higher accuracy extractions and deals with ambiguous extractions, although at the cost of requiring human effort to label training data. The results demonstrate that reference-set based extraction outperforms the current state-of-the-art systems that rely on structural or grammatical clues, which is not appropriate for unstructured, ungrammatical text. Even the fully automatic case, which constructs its own reference set for automatic extraction, is competitive with the current state-of-the-art techniques that require labeled data. Reference-set based extraction from unstructured, ungrammatical text allows for a whole category of sources to be queried, allowing for their inclusion in data integration systems that were previously limited to structured and semi-structured sources.

PROBLEM DEFINITION

As more and more information comes online, the ability to process and understand this information becomes more and more crucial. Data integration attacks this problem by letting users query heterogeneous data sources within a unified query framework, combining the results to ease understanding. However, while data integration can integrate data from structured sources such as databases, semi-structured sources such as that extracted from Web pages, and even Web Services, this leaves out a large class of useful information: unstructured and ungrammatical data sources. We identify such unstructured, ungrammatical data as “posts". Posts range in source from classified ads, auction listings, and forum postings to blog titles or paper references. The goal of this project is to structure sources of posts, such that they can be queried and included in data integration systems.

PROPOSED METHOD

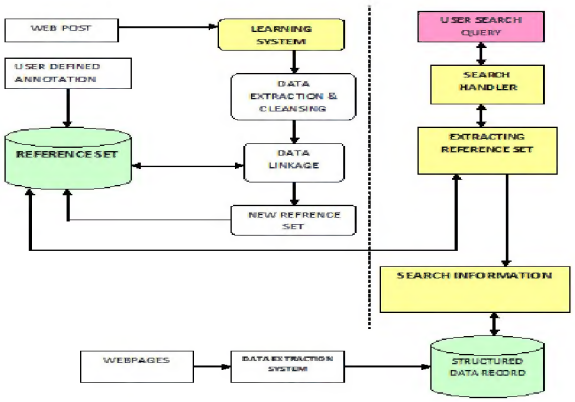

Figure 1: System Architecture

To design Web-Scale Information Extraction Using Wrapper Induction Approach for an unstructured web data the following systems will be considered:

- Learning System
- Data Extraction System
- User Search Query System

Learning System

A learning system is responsible for learning a new set of extraction rules for specific sites. A single web site may contain pages conforming to multiple different templates, from each website all samples of pages are collected and are clustered using Shingle based signature which is computed for each web page based on html tags.

Data Extraction System

An Extraction system, the learnt rules are applied to the stream of crawled wed pages to extract records from them. For each incoming web page, the shingle based signature and page URL are used to find the matching rule for the page, which is then applied to extract the record for the page.

User Search Query System

A Search Query System, are used to search matching records based user query. For each request query will be matched based on the rules of learning system. The key contribution of the project is for information extraction that exploits reference sets, rather than grammar or structure based techniques. The project includes the following contributions:

- An automatic learning system for matching and extraction of reference set.
- A method that selects the appropriate reference sets from a repository and uses them for extraction and annotation without training data.
- An automatic method for constructing reference sets from the posts themselves.
- An automatic method for web post record extraction using reference set for searching accurate information.

The proposed approach will be implemented by collection of known entities and their attributes, which refer as "reference set," A reference set can be constructed from structured sources, such as databases, or scraped from semi-structured sources such as collections of Web pages. A reference set can even be constructed automatically from the unstructured, ungrammatical text itself. It follows the following methodology for information extraction from unstructured, ungrammatical data sources:

- Automatically Choosing the Reference Sets
- Matching Posts to the Reference Set
- Extraction using reference sets
- Automatically Constructing Reference Sets for Extraction

- A Learning Approach to Reference-Set Based Extraction
- Extracting Data from unstructured data sources

CONCLUSIONS

Thus we proposed a system based on Machine Learning technique to obtain the structured data records from different unstructured and non-template based websites. The proposed approach will be implemented by collection of known entities and their attributes, which refer as "*reference set*," A reference set can be constructed from structured sources, such as databases, or scraped from semi-structured sources such as collections of Web pages. A reference set can even be constructed automatically from the unstructured, ungrammatical text itself. Thus this project implements methods to exploit reference sets for extraction using machine learning techniques. The machine learning approach provides higher accuracy extractions and deals with ambiguous extractions, although at the cost of requiring human effort to label training data.

REFERENCES

1. Gulhane, P.; Madaan, A.; Mehta, R.; Ramamirtham, J.; Rastogi, R.; Satpal, S.; Sengamedu, S.H.; Tengli, A.; Tiwari, C.; Web-scale information extraction with vertex. Data Engineering (ICDE), 2011 IEEE 27th International Conference on Digital Object Identifier Publication Year: 2011, Page(s): 1209 - 1220
2. Laender, A. H. F., Ribeiro-Neto, B., DA Silva and Teixeira, A brief survey of Web data extraction tools. SIGMOD Record 31(2): 84-93, 2002.
3. Wei Liu; Xiaofeng Meng; Weiyi Meng; ViDE: A Vision-Based Approach for Deep Web Data Extraction Knowledge and Data Engineering, IEEE Transactions on Volume: 22 Publication Year:2010, Page(s): 447 – 460
4. Ril off, E., Automatically constructing a dictionary for information extraction tasks. Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93), pp. 811-816, AAAI Press/ The MIT Press, 1993.
5. Chang, C.-H., Hsu, C.-N., and Lui, S.-C. Automatic information extraction from semi-Structured Web Pages by pattern discovery. Decision Support Systems Journal, 35(1): 129-147, 2003.
6. Arocena, G. O. and Mendelzon, A. O., Web OQL: Restructuring documents, databases, and Webs. Proceedings of the 14th IEEE International Conference on Data Engineering (ICDE), Orlando, Florida, pp. 24-33, 1998.
7. Hammer, J., McHugh, J. and Garcia-Molina, Semi structured data: the TSIMMIS experience. In Proceedings of the 1st East-European Symposium on Advances in Databases and Information Systems (ADBIS), St. Petersburg, Rusia, pp. 1-8, 1997.
8. Saiiuguet, A. and Azavant, F., Building intelligent Web applications using lightweight wrappers. Data and Knowledge Engineering 36(3): 283-316, 2001.